Query Based Tri-Clustering (QBTC)

N. Narmadha
Department of Computer Science
Periyar University
Salem, India
mahanarmadha@gmail.com

***Abstract*- In the world of big data, accumulated 3D Gene Expression Data is increased rapidly, therefore a novel query based tricluster is proposed in this work to extract maximum similarity tricluster from the given 3D data. The main advantage of this proposed work is query tricluster is the identification of customized tricluster with respect to the given query. This query is the most valuable or functionable gene. The performance of the proposed work is studied with the stimulated data. It has observed the query Tri performance well in extracting constant, shifting and scaling pattern tricluster.**

Keywords- Query, Triclustering, 3D Gene Expression Data, Tricluster, Similarity Measure

1. INTRODUCTION

Owe to recent development in the field of microarray technology, more quantity of gene-sample-time microarray data (GST) or three dimensional microarray data are generated very easily and frequently for analysis. It contains the expression levels of a group of genes under a set of samples/conditions during a series of time points (Schuh, 2004). Three dimensional (3D) Microarray dataset is a dataset contains 3 types of variables (gene, sample, and time point). In general, each cell m_{ijk} in a 3D dataset represents the value of i^{th} row under j^{th} column at k^{th} time space. It can also be viewed as a two-dimensional matrix, such that each cell $m_{i,j}$ contains the time series with respect to i^{th} row under j^{th} column (J. Bagyamani, 2013).

In this work, a novel triclustering algorithm is developed to extract maximum similar tricluster from the 3D gene expression data with respect to given query gene. Here, query gene acts as a seed in extracting the tricluster. A similarity score between two genes is defined as in (Faris Alqadah Joel S. Bader, 2018) and similarity score for tricluster is defined. To our knowledge this is the first time that query gene based similarity score used for tricluster problem solving. Empirical study is conducted on simulated data shows that proposed triclustering algorithm performs well in extracting maximum similar tricluster with respect to query gene (J.Bagyamani D. K., 2011).

This paper is organized as follows: Section 2 describes the Literature Review/Related work needed for this research work. Proposed work is described in section 3. Section 4 elaborates the experiment analysis. Finally, section 5 concludes the proposed work with possible future enhancement.

2. RELATED WORK

This section is to provide the general overview of related works in the field of 3D microarray gene expression data analysis. In particular, for those works related to the query based clustering and biclustering technique are listed below.

The main of the algorithm is to extract maximum similarity bicluster after applying feature selection using multiple node deletion (Xiaowen Liu, 2006).To introduce QDB, a novel Bayesian query-driven biclustering framework to guide the pattern search. A resolution sweep approach that successfully grows the enriched biclusters from small sets of seed genes. In that modularity of the biclusters is established and the relevant conditions are identified. At last it mainly focused on the missing values naturally and it performs well on artificial data from the biclustering benchmark study (J.Bagyamani D. K., 2010).

To propose generalize query-based biclustering for high dimensional data. The framework performs a local approach for query-based biclustering. Thus the local approach gives the exact bicluster; it also provides the higher quality when compared to the other query based methods and the QBBC(Query Based BiClustering) is efficient and scalable (Tao JIANG).

SIMBIC and SIMBIC+ are the two Biclustering Models is used to extract the bicluster based on the query gene and condition and includes the query input also (Rui Henriques, 2018). The biclustering with coherent bicluster and Constant bicluster with reference to the query gene are extracted (D. Gutiérrez-Avilés, 2011).

The latest query-based biclustering algorithms like ISA and QDB, the mainly focused on the bicluster quality and the outcome against noisy seed sets and biological relevance. But the ProBic's is successfully retrieving the biologically relevant data with high quality biclusters that retain their seed genes and it main aims to handle noisy seeds (J. Bagyamani, 2013).

The OPSM query is difficult to handle real life exploratory data analysis processing and it hard to capture subjective interestingness aspects. OPSM query method introduces two constrained that is user defined constraints based on their query (Rui Henriques, 2018). The experiments are tested on real datasets and the experiment results shown the multi-dimension index (cIndex) and the enumerating sequence index (esIndex) based on their queries and it provides the better performance than brute force search (J.Bagyamani D. K., 2011).

GO-Cluster uses the tree structure of the Gene Ontology database it deals with the numerical cluster algorithms and it is mainly focused on gene expression data. In that the expected correlation between genetic co-regulation and a common biological process is not needed. Thus the visualization of gene expression data shows the various levels of the ontology tree (Thomas Dhollander, 2007).

3. PROPOSED WORK

3.1. Problem Statement

Given an n×m×t similarity matrix S(I,J.K), the maximum similarity tricluster problem (TriMSB) is to find a tricluster S(I′,J′.K′) with I′⊆I , J⊆J′ and K⊆K′ such that S(I′,J′.K′) is maximized. The tricluster S(I′,J′.K′) is called the maximum similarity tricluster of S(I,J.K) (Xiaowen Liu, 2006).

3.2. Coherent Tricluster

Genes involved in common processes are often co-expressed. In this paper, constant tricluster with reference to the query gene and coherent tricluster with reference to the query gene are extracted. The coherent additive tricluster and coherent multiplicative tricluster is represented.

3.3. Similarity Score between Genes at Time point

An element a_{ij} of expression matrix is represented as A(I,J) and a reference gene i*∈ I, it is defined as $d_{ij} = |a_{ij} - a_{i*j}|$.To ignore the elements with big $d_{ij}$for that set a threshold $\alpha . d_{avg}$ is shown in equation (1)

$$d_{avg} = \frac{\sum_{i \in I \& j \in J} d_{ij}}{|I||J|} \quad (1)$$

The average distance value of all elements in A(I,J). If $d_{ij} > \alpha . d_{avg}$the two elements a_{ij} and a_{i*j} are not a similarity and set the similarity s_{ij} to be 0. Otherwise similarity score is shown in equation (2)

$$1 - \frac{d_{ij}}{\alpha . d_{avg}} + \beta \quad (2)$$

β is the bonus for small d_{ij} and it is used as the increase of the similarity score for small d_{ij} and ignore d_{ij}′s greater than the threshold.

$$s_{ij} = \begin{cases} 0 & if\ d_{ij} > \alpha . d_{avg} \\ 1 - \frac{d_{ij}}{\alpha . d_{avg}} + \beta & otherwise \end{cases} \quad (3)$$

(Xiaowen Liu, 2006) When $d_{ij} > \alpha . d_{avg}$ to have $\frac{d_{ij}}{\alpha . d_{avg}} \leq 1$, where $s_{ij}$is always ≥ 0. Here S(I,J) is denoted as the n×m similarity matrix includes the set of rows I and the set of columns J with every element s_{ij} that can be shown in equation (3)

3.4. Maximum Similarity Score for Tricluster (TriMSB)

3.4.1. Definition

Given an n×m×t similarity matrix S(I,J.K), the maximum similarity tricluster problem (Tri_{MSB}) is to find a tricluster S(I′,J′.K′) with I′⊆I , J⊆J′ and K⊆K′ such that S(I′,J′.K′) is maximized. The tricluster S(I′,J′.K′) is called the maximum similarity tricluster of S(I, J, K) (Rubio-Escudero, 2014).

3.5. Algorithm for Maximum similarity Score for Triclustering

Input

1. Gene Expression Matrix A(I,J,K)
2. Reference Gene i*

Output A Maximum Similarity Tricluster

Steps

1. Compute Similarity Matrix S(I,J,K) using Query for reference gene i*
2. Computation additive tricluster similarity score
3. smat- similarity matrix at each time point
4. S-similarity score for bicluster at each time point
5. tval-similarity score for tricluster
6. Find the Multiplicative Tricluster similarity score using (3, 4, 5)
7. End while maximum similarity score for additive tricluster and multiplicative tricluster is obtain

Table: 1 List of various Measures for Triclustering

Measures	Descriptions and its Formula
MCV	$\sum_m \sum_n (A_{mn}-\bar{A})(B_{mn}-\bar{B}) \quad / \sqrt{\left(\sum_m \sum_n (A_{mn}-\bar{A})^2\right)\left(\sum_m \sum_n (B_{mn}-\bar{B})^2\right)}$ Where $\bar{A}=\frac{\sum m \sum n(A_{mn})}{m*n}$, $\bar{B}=\frac{\sum m \sum n(B_{mn})}{m*n}$ The range of MCV is [0 ,1]
MSR_{3D}	$MSR_{3D}(TC)=\frac{\sum_{g\in G,c\in C,t\in T} r^2 gct}{\#G*\#C*\#T}$ Where r_{gct} can be defined as $r_{gct}=TC_V(g,c,t)+M_{CT}(g)+M_{GT}(c)+M_{GC}(T)-M_G(c,t)-M_C(g,t)-M_T(g,c)-M_{GCT}$
Tri_{MSB}	$S(I',J'.K')=\min\{\min_{i\in r} s(i,j,K'), \min_{j\in r} s(i,j,K')\}$

Table 1 shows the list of various correlation measures for triclustering it includes MCV, MSR3D, and TriMSB, these are the homogeneity measure it is used to evaluate the quality of the tricluster which contain three dimensional data such as gene, sample and time point.

4. EXPERIMENTAL ANALYSIS

4.1. Dataset Generation

In detail, the matrix with implanted constant tricluster is generated with four steps:

1) Generate a 50×10×5 matrix A such that all elements of A are 1's

2) Implant the 11× 10 matrix into the A without overlap such that all elements are values.

3) Replace δ into 50×10×5 matrix A with random noise.

4) For each test on constant, additive, and multiplicative to generate the tricluster based on the user query.

4.2. Parameter Selection

The experiment results are performed some simulations on selecting the parameters.

Table: 2 Parameter Setting for Tricluster

Parameter Notation	values
Alpha	$\alpha \in [0.2,0.4]$
beta	$\beta \in [0.0,0.5]$
gamma	$\Upsilon \in [\beta+0.7, \beta+0.9]$
Noise	$\delta \in [0,0.25]$

Table 2 describes the Parameter setting for Tricluster. The performance of MCV, MSR3D, and TriMSB, when compared these measures TriMSB gives better performance than MCV, MSR3D. This algorithm is developed in Matlab Toolbox R2013b. Thus the results are shown in the table 3. Table 4 describes the evaluation of various type of query with 3D data.

Table: 3 Performance of MCV, MSR3D , and TriMSB

Tricluster	Constant	Additive	Multiplicative
MCV	0	0	0
MSR_{3D}	0	0	0
TriMSB	1	1	1

Table: 4 Evaluation of various type of Query and its Tricluster type

Query Type	Size of Embedded Tricluster			Size of Extracted Maximum similarity Tricluster			No. of Iteration	Tricluster Type
Constant	50	10	5	49	10	2	63	Constant
Additive	50	10	5	47	10	4	60	Additive
Multiplicative	50	10	5	48	10	3	62	Multiplicative

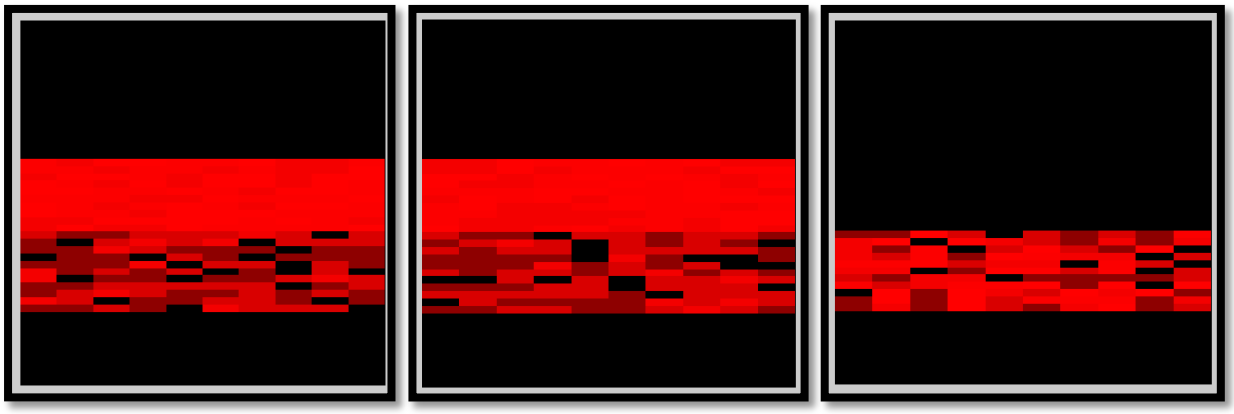

Figure 1. Heatmap Representation of constant, Additive and Multiplicative Tricluster

Figure 1 shows the representation of the heatmap for the constant, Additive and Multiplicative Tricluster it shows the quality of the tricluster.

5. CONCLUSION

The novel measure is proposed for quality of tricluster among genes, sample and time point. Triclustering algorithm is developed to extract maximum similar tricluster from the 3D gene expression data with respect to given query gene. It works better to extract the maximum similar tricluster for the constant, shifting and scaling pattern tricluster.

6. ACKNOWLEDGEMENT

The first author acknowledges the UGC for the financial support to her research under the UGC NFSC Scheme (Student Id: 29799/(SC -2017)).

REFERENCES

D. Gutiérrez-Avilés, C. R.-E. (2011). Triclustering on temporary microarray data using the TriGen algorithm. 11th International Conference on Intelligent Systems Design and Applications.

Faris Alqadah Joel S. Bader, R. A. (2018). Query-based Biclustering using Formal Concept Analysis. Johns Hopkins University,Baltimore .

Hui Zhao, L. C. (2011). Query-based biclustering of gene expression data using Probabilistic Relational Models. BMC Bioinformatics.

J. Bagyamani, D. K. (2013, Mar). Comparsion of Biological Significance of Biclusters of SIMBIC and SIMBIC+ Biclustering Models. ACEEE Int.J.on Information Technology, 3.

J.Bagyamani, D. K. (2010). SIMBIC :SIMilarity Based BIClustering of Expression Data. Springer-Verlag Berlin Heidelberg.

J.Bagyamani, D. K. (2011, Feb). Biological Significance of Gene Expression Data Using Similarity based Biclustering Algorithm. International Journal of Biometrics and Bioinformatics (IJBB), 4(6).

Rubio-Escudero, D. G.-A. (2014). Mining 3D Patterns from Gene Expression Temporal Data: A New Tricluster Evaluation Measure. The Scientific World Journal, 16.

Rui Henriques, S. C. (2018). Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey. ACM Computing Surveys.

Schuh, B. A. (2004). Gene-Ontology-based clustering of gene expression data. Bioinformatics , Oxford University Press, 20(16).

Tao JIANG, Z. L.. Constrained query of order-preserving submatrix in gene expression data. School of Computer Science and Technology, Northwestern Polytechnical University .

Thomas Dhollander, Q. S. (2007). Query-driven module discovery in microarray data. Published by Oxford University Press.

Xiaowen Liu, L. W. (2006, Nov). Computing the maximum similarity bi-clusters of gene expression data. Bioinformatics.